2. Sager MA, Franke T, Inouye SK *et al.* Functional outcomes of acute medical illness and hospitalization in older persons. Arch Intern Med 1996; 156: 645–52.
3. Cornette P, Swine C, Malhomme B, Gillet JB, Meert P, D'Hoore W. Early evaluation of the risk of functional decline following hospitalization of older patients: development of a predictive tool. Eur J Public Health 2006; 16: 203–8.
4. Thomas R. Focus on functional decline in hospitalized older patients. J Gerontol A Biol Sci Med Sci 2002; 57A: M567–8.
5. Graf C. Functional decline in hospitalized older patients. Am J Nurs 2006; 106: 58–67.
6. Inouye SK, Wagner R, Acampora D *et al.* A predictive index for functional decline in hospitalized elderly medical patients. J Gen Intern Med 1993; 8: 645–52.
7. Hébert R. Functional decline in old age. Can Med Assoc J 1997; 157: 1037–45.
8. Stuck A, Walthert J, Nikolaus T, Büla CJ, Hohmann C, Beck JC. Risk factors for functional decline in community-living elderly people: a systematic review. Soc Sci Med 1999; 48: 445–69.
9. McCusker J, Kakuma R, Abrahamowicz M. Predictors of functional decline in hospitalized elderly patients: a systematic review. J Gerontol A Biol Sci Med Sci 2002; 57A: M569–77.
10. Meldon SW, Mion LC, Palmer RM *et al.* A brief risk-stratification tool to predict repeat emergency department visits and hospitalizations in older patients discharged from the emergency department. Acad Emerg Med 2003; 10: 224–32.
11. Hustey FM, Mion LC, Connor JT, Emerman CL, Campbell J, Palmer RM. A brief risk stratification tool to predict functional decline in older adults discharged from emergency departments. J Am Geriatr Soc 2007; 55: 1269–74.
12. McCusker J, Bellavance F, Cardin S, Trépanier S, Verdon J, Ardman O. Detection of older people at increased risk of adverse health outcomes after an emergency visit: the ISAR screening tool. J Am Geriatr Soc 1999; 47: 1229–37.
13. Dendukuri N, McCusker J, Belzile E. The identification of seniors at risk screening tool: further evidence of concurrent and predictive validity. J Am Geriatr Soc 2004; 52: 290–6.
14. Vandewoude MF, Geerts CA, d'Hooghe AH, Paridaens KM. A screening tool to identify older people at risk of adverse health outcomes at the time of hospital admission. Tijdschr Gerontol Geriatr 2006; 37: 203–9.
15. Geyskens K, De Ridder K, Sabbe M *et al.* Prediction of functional decline in elderly patients discharged from the accident and emergency department. Tijdschr Gerontol Geriatr 2008; 39: 16–25.
16. Moons P, De Ridder K, Geyskens K *et al.* Screening for risk of readmission of patients aged 65 years and above after discharge from the emergency department: predictive value of four instruments. Eur J Emerg Med 2007; 14: 315–23.
17. Kenis C, Geeraerts A, Braes T, Milisen K, Flamaing J, Wildiers H. The Flemish version of the Triage Risk Screening Tool (TRST): a multidimensional short screening tool for the assessment of elderly patients. Crit Rev Oncol Hematol 2006; 60(Suppl 1): 31.
18. Katz S, Akpom CA. Index of ADL. Med Care 1976; 14: 116–8.
19. Ellis G, Langhorne P. Comprehensive geriatric assessment for older hospital patients. Br Med Bull 2005; 71: 45–59.

doi: 10.1093/ageing/afp097
Published electronically 25 June 2009

Inter-rater reliability of STOPP (Screening Tool of Older Persons' Prescriptions) and START (Screening Tool to Alert doctors to Right Treatment) criteria amongst physicians in six European countries

SIR—Inappropriate prescribing (IP) encompasses the use of medicines where the risk of an adverse drug event (ADE) outweighs the clinical benefit, particularly when safer or more effective alternatives are available [1, 2]. IP also includes the use of medicines that increase the likelihood of drug–drug and drug–disease interactions, the mis-prescribing of medicines (incorrect dose, frequency and duration) and the under-use of clinically indicated medicines [3–5]. IP is highly prevalent in older people and has been associated with preventable ADEs, hospitalisation, institutionalisation, death and resource wastage [6–12]. With increasing proportions of older people worldwide, quality and safety of prescribing are becoming a global healthcare concern [5, 13].

One way of identifying IP is to use prescribing indicators such as the recently validated STOPP (Screening Tool of Older Persons' Prescriptions) and START (Screening Tool to Alert doctors to Right Treatment) criteria [14]. STOPP comprises 65 indicators for potentially inappropriate prescribing including drug–drug and drug–disease interactions, therapeutic duplication and drugs that increase the risks of cognitive decline and falls (Appendix 1 in the supplementary data at *Age and Ageing* online) [14]. START incorporates 22 evidence-based indicators for prescribing omissions in older people (Appendix 2 in the supplementary data at *Age and Ageing* online) [14]. STOPP/START criteria are organised according to physiological systems for ease of use. Their content validity was established by a Delphi consensus process in which 18 experts in geriatric pharmacotherapy from Ireland and the United Kingdom participated [14]. A recent study showed that 35% of 715 acutely ill older patients requiring hospitalisation were regularly prescribed at least one potentially inappropriate medication according to STOPP criteria and 12% of admissions were directly attributable to associated serious ADEs [15]. Another study of 600 older patients showed that 58% were not prescribed clinically indicated medications without contraindication according to START criteria [16].

Prospective randomised controlled trials are needed to test whether routine clinical application of STOPP/START criteria can significantly improve prescribing appropriateness and reduce drug-related morbidity. However, before demonstrating effects on patient outcome, a screening tool must be generalisable and reliable. Inter-rater reliability of STOPP/START criteria was substantial when tested between two researchers (kappa coefficient 0.75 STOPP criteria and 0.68 START criteria) [14]. Further evaluation of reliability between health professionals practicing in different countries is warranted to determine if STOPP/START criteria are generalisable. Accordingly, the aim of this study was to

determine the inter-rater reliability of STOPP and START criteria between multiple physicians practicing in different European centres.

Methods

Twenty datasets were selected from a cohort of 200 consecutive patients aged ≥65 years admitted acutely to the general medical services of a university teaching hospital in Ireland who were participating in a concurrent randomised controlled trial designed to evaluate the effect of an intervention on prescribing appropriateness. The 20 datasets were specifically selected to represent patients with complex comorbidities and an appreciable incidence of potentially inappropriate prescriptions according to STOPP/START criteria. Ethical approval was obtained for anonymous use of these datasets in this inter-rater reliability study.

Each dataset was compiled from chart review and patient and/or carer interview at the time of hospitalisation, with supplementary information on diagnoses and prescriptions being sought from the patient's general practitioner and/or community pharmacist when required. Datasets comprised age, gender, current and past diagnoses, detailed medication history, drug allergies, blood pressure profile, electrocardiograph results, serum biochemistry, glucose, lipid profile, urinalysis and estimated creatinine clearance using the Cockcroft–Gault equation [17], chosen instead of the Modified Diet in Renal Disease formula [18–20] as the latter is not well validated in patients aged >70 years [20–22].

The 20 patients' mean (±SD) age was 76.8 (±5.4) years, and 50% were female. The total number of prescribed medications was 181, median 9, interquartile range 6–11. Two of the STOPP/START originators (PG, DO'M) from the coordinating centre in Ireland discussed the 20 datasets in detail and reached complete agreement in terms of prescribing appropriateness according to STOPP/START criteria. This combined level of agreement (labelled 'rater 1') was set as the standard against which other physicians' ratings would be compared. Nineteen datasets had at least one potentially inappropriate prescription according to STOPP criteria (median 2; range 0–5). Eleven datasets had at least one prescribing omission according to START criteria (median 1; range 0–4).

Eight hospital physicians (labelled 'raters 2–9') with no prior experience of using STOPP/START criteria participated in the study. These physicians were based in teaching hospital geriatric medicine units in Belgium ($n = 1$), Czech Republic ($n = 2$), Italy ($n = 3$), Spain ($n = 1$) and Switzerland ($n = 1$). STOPP/START criteria were translated from English into Czech, French, Italian and Spanish (available on request from the corresponding author) to facilitate local application of the criteria. The 20 datasets were also translated. A teleconference facilitated by the coordinating centre in Ireland afforded all raters the opportunity to resolve any difficulties with translation or interpretation of the criteria before application. Two criteria were clarified: (i) therapeutic and maintenance doses of proton pump inhibitors (STOPP C4) and (ii) inclusion criteria for drug-class duplication (STOPP J1). All physicians then independently assessed the incidence of 65 STOPP and 22 START criteria in each of the 20 datasets and were invited to give written comments if necessary.

Responses of raters 2–9 were cross-tabulated with those of rater 1. Inter-group responses between physicians from Italy and the Czech Republic were also evaluated, to determine reliability independent of the STOPP/START originators. Data were analysed using SPSS 15.0 for Windows (SPSS Inc., Chicago, IL, USA). Inter-rater reliability analysis using the kappa statistic (chance corrected measure of agreement) was performed to determine consistency between raters [23]. The kappa statistic was interpreted as poor if ≤0.2, fair if 0.21–0.40, moderate if 0.51–0.6, substantial if 0.61–0.8 and good if 0.81–1.00 [23]. Proportions of positive and negative agreements were calculated [24]. Written comments were analysed to determine whether disagreements were due to misjudgement of appropriateness, difficulty with criteria or case interpretation.

Table 1. Inter-rater reliability of STOPP and START criteria between 9 hospital physicians on 20 datasets with 181 medications

Rater combination	A	B	C	D	Ppos	Pneg	Kappa (95% CI)
STOPP criteria							
Rater 1 * rater 2	1,255	4	0	41	0.99	0.95	0.95 (0.91–0.99)
Rater 1 * rater 3	1,254	5	3	38	0.99	0.90	0.90 (0.83–0.97)
Rater 1 * rater 4	1,254	5	3	38	0.99	0.90	0.90 (0.83–0.99)
Rater 1 * rater 5	1,255	4	0	41	0.99	0.95	0.95 (0.91–0.99)
Rater 1 * rater 6	1,258	1	2	39	0.99	0.96	0.96 (0.92–1)
Rater 1 * rater 7	1,257	2	1	40	0.99	0.96	0.96 (0.92–1)
Rater 1 * rater 8	1,253	6	3	38	0.99	0.89	0.89 (0.82–0.96)
Rater 1 * rater 9	1,250	9	0	41	0.99	0.90	0.90 (0.83–0.96)
Median (IQR)					0.99	0.93	0.93 (0.90–0.96)
START criteria							
Rater 1 * rater 2	417	3	2	18	0.99	0.88	0.87 (0.76–0.98)
Rater 1 * rater 3	417	3	3	17	0.99	0.85	0.84 (0.72–0.97)
Rater 1 * rater 4	418	2	1	19	0.99	0.92	0.92 (0.84–1)
Rater 1 * rater 5	417	3	0	20	0.99	0.93	0.93 (0.84–1)
Rater 1 * rater 6	416	4	3	17	0.99	0.83	0.82 (0.69–0.95)
Rater 1 * rater 7	415	5	5	15	0.98	0.75	0.74 (0.58–0.89)
Rater 1 * rater 8	413	7	1	19	0.99	0.83	0.82 (0.69–0.94)
Rater 1 * rater 9	414	6	0	20	0.99	0.87	0.86 (0.75–0.97)
Median (IQR)					0.99	0.86	0.85 (0.82–0.91)

A, both raters agreed criterion not fulfilled; B, rater 1 scored criterion not fulfilled and rater 2 scored criterion as being fulfilled; C, rater 1 scored criterion as fulfilled and rater 2 scored criterion as not fulfilled; D, both raters scored criterion as being fulfilled; ppos, proportion of positive agreement; pneg, proportion of negative agreement; CI, confidence interval; IQR, interquartile range.

Results

Columns A, B, C and D in Tables 1 and 2 indicate the status of agreement between raters. For example, raters 1 and

Table 2. Inter-rater reliability of STOPP and START between three physicians from Italy (raters 3, 4 and 5) and two physicians from the Czech Republic (raters 6 and 7) on 20 datasets with 181 medications

Rater combination	A	B	C	D	Ppos	Pneg	Kappa (95% CI)
STOPP criteria							
Rater 3 * rater 4	1,257	0	0	43	1.00	1.00	1.00 (NA)
Rater 3 * rater 5	1,253	4	2	41	0.99	0.93	0.93 (0.87–0.99)
Rater 4 * rater 5	1,253	4	2	41	0.99	0.93	0.93 (0.87–0.99)
Rater 6 * rater 7	1,257	3	1	39	0.99	0.95	0.95 (0.90–0.99)
START criteria							
Rater 3 * rater 4	418	2	1	19	0.99	0.93	0.92 (0.84–1.00)
Rater 3 * rater 5	415	5	2	18	0.99	0.84	0.83 (0.71–0.95)
Rater 4 * rater 5	416	3	1	20	0.99	0.91	0.90 (0.81–0.99)
Rater 6 * rater 7	413	4	7	16	0.99	0.74	0.73 (0.58–0.88)

A, both raters agreed criterion not fulfilled; B, rater 1 scored criterion not fulfilled and rater 2 scored criterion as being fulfilled; C, rater 1 scored criterion as fulfilled and rater 2 scored criterion as not fulfilled; D, both raters scored criterion as being fulfilled; ppos, proportion of positive agreement; pneg, proportion of negative agreement; CI, confidence interval; NA, not applicable.

2 agreed that STOPP criteria were not identified in 1,255 instances (column A). In four instances, rater 1 did not identify a STOPP criterion, but rater 2 did (column B). There were no instances where rater 2 identified a STOPP criterion that rater 1 did not (column C). In 41 instances, both raters 1 and 2 identified a STOPP criterion (column D). The median (IQR) kappa coefficient between raters was 0.93 (0.90–0.96) for STOPP criteria and 0.85 (0.82–0.91) for START criteria.

Disagreement occurred with six STOPP criteria. Two raters disagreed that aspirin was potentially inappropriate without coronary, cerebral or peripheral arterial occlusive symptoms or risk factors (STOPP A13). Three raters misclassified second-generation antihistamines as first-generation antihistamines (STOPP B13 and H3). Loop diuretics were prescribed without indication in four cases; however, six raters inferred an indication of ankle oedema or hypertension, thereby resulting in disagreement over STOPP criteria A2 and A3. A typical neuroleptic was deemed inappropriate by one rater (STOPP B8), although it was prescribed for behavioural and psychological symptoms of dementia.

One rater judged it appropriate to omit a statin in one patient with coronary artery disease (START A5) and severe heart failure, as it was 'unlikely to alter outcome'. Four raters did not apply this criterion, as they 'did not have sufficient clinical information to measure life expectancy'. Two raters wanted more clinical details before applying the START criterion C2 (anti-depressant with depressive symptoms lasting >3 months) and START criterion E1 (disease modifying anti-rheumatic drug with moderate–severe rheumatoid arthritis). Five raters identified the START criterion D2 (fibre supplement for chronic symptomatic diverticular disease with constipation) in two patients with chronic constipation, though diverticular disease was not specifically documented. Medications without indication were sometimes interpreted as indicating an underlying disease for which another medication was omitted e.g. one rater interpreted nitrate use as indicating underlying angina and recommended a beta-blocker (START A8).

Discussion

Inter-rater reliability of STOPP/START criteria is good when tested between multiple physicians across six European centres. The more comprehensive clinical and medication details used in this study are likely to account for the higher level of inter-rater reliability than reported previously [14]. Disagreements in a minority of instances reflected the fact that details on functional status and life expectancy were not provided with the cases, though clearly, these are important considerations when applying STOPP/START criteria. Differences in prescribing guidelines and formularies between countries could cause disagreement with STOPP/START criteria, but no physician in this study reported this.

A high level of familiarity is required to efficiently apply 'pencil and paper' versions of STOPP/START criteria in clinical practice. This reality emphasises the need for computerised automation of STOPP/START whereby linkage of specific diseases or symptoms with specific medicines would lead to rapid identification of potentially inappropriate prescriptions according to STOPP criteria and omission of indicated drugs according to START criteria. The good inter-rater reliability demonstrated by this study indicates that results of studies on the prevalence of potentially inappropriate prescribing identified by STOPP/START criteria are comparable between countries.

Key points

- Inter-rater reliability of STOPP and START criteria is good when tested between multiple physicians practicing independently in different European centres.
- STOPP and START criteria are generalisable across different European countries and languages.

Conflicts of interest

No conflicts of interest.

Funding

Health Research Board of Ireland (Clinical Research Training Fellowship CRT/2006/029) and the Czech Ministry of Health Internal Grant Agency (Grant Number 10029-4).

Supplementary data

Supplementary data are available online at *Age and Ageing* online.

PAUL GALLAGHER[1,*], JEAN-PIERRE BAEYENS[2], EVA TOPINKOVA[3], PAVLA MADLOVA[3], ANTONIO CHERUBINI[4], BEATRICE GASPERINI[4], ALFONSO CRUZ-JENTOFT[5], BEATRIZ MONTERO[5], PIERRE OLIVIER LANG[6], JEAN-PIERRE MICHEL[6], DENIS O'MAHONY[1]

[1]*Department of Geriatric Medicine, Cork University Hospital, Wilton, Cork, Ireland*

[2]*Department of Geriatric Medicine, AZ Damiaan Oostende, Oostende, Belgium*

[3]*Department of Geriatric Medicine, First Faculty of Medicine, Charles University, Prague, Czech Republic*

[4]*Department of Clinical and Experimental Medicine, Institute of Gerontology and Geriatrics, University of Perugia Medical School, Perugia, Italy*

[5]*Servicio de Geriatria, Hospital Universitario Ramón y Cajal, Madrid, Spain*

[6]*Rehabilitation and Geriatric Department, Geneva Medical School and University Hospitals, Geneva, Switzerland*

Email: pfgallagher77@eircom.net

[*]*To whom correspondence should be addressed*

References

1. Beers MH, Ouslander JG, Rollinger I, Brooks J, Reuben D, Beck JC. Explicit criteria for determining inappropriate medication use in nursing homes. Arch Intern Med 1991; 151: 1825–32.
2. Beers MH. Explicit criteria for determining potentially inappropriate medication use by the elderly. Results of a US consensus panel of experts. Arch Intern Med 1997; 163: 2716–4.
3. Hanlon JT, Schmader KE, Ruby CM, Weinberger M. Suboptimal prescribing in older inpatients and outpatients. J Am Geriatr Soc 2001; 49: 200–9.
4. Simonson W, Feinberg JL. Medication-related problems in the elderly: defining the issues and identifying solutions. Drugs Aging 2005; 22: 559–69.
5. Spinewine A, Schmader KE, Barber N *et al.* Appropriate prescribing in elderly people: how well can it be measured and optimised? Lancet 2007; 370: 173–84.
6. Lindley CM, Tully MP, Paramsothy V, Tallis RC. Inappropriate medication is a major cause of adverse drug reactions in elderly patients. Age Ageing 1992; 21: 294–300.
7. Gallagher P, Barry P, Ryan C, Hartigan I, O'Mahony D. Inappropriate prescribing in an acutely ill population of elderly patients as determined by Beers' criteria. Age Ageing 2008; 37: 96–101.
8. Klarin I, Wimo A, Fastbom J. The association of inappropriate drug use with hospitalisation and mortality: a population based study of the very old. Drugs Aging 2005; 22: 69–82.
9. Lau DT, Kasper JD, Potter DE, Lyles A, Bennett RG. Hospitalization and death associated with potentially inappropriate medication prescriptions among elderly nursing home residents. Arch Intern Med 2005; 165: 68–74.
10. Gurwitz, JH, Field TS, Avorn J *et al.* Incidence and preventability of adverse drug events in nursing homes. Am J Med 2000; 109: 87–94.
11. Zuckerman IH, Langenberg P, Baumgarten M *et al.* Inappropriate drug use and risk of transition to nursing homes among community-dwelling older adults. Med Care 2006; 44: 722–30.
12. Fick DM, Waller JL, Maclean JR *et al.* Potentially inappropriate medication use in a Medicare managed care population: association with higher costs and utilization. J Manag Care Pharm 2001; 7: 407–13.
13. Laroche ML, Charmes JP, Boutheir F, Merle L. Inappropriate medications in the Elderly. Clin Pharmacol Ther 2009; 85: 94–7.
14. Gallagher P, Ryan C, Byrne S, Kennedy J, O'Mahony D. STOPP (Screening Tool of Older Persons' Prescriptions) and START (Screening Tool to Alert Doctors to Right Treatment): consensus validation. Int J Clin Pharmacol Ther 2008; 46: 72–83.
15. Gallagher P, O'Mahony D. STOPP (Screening Tool of Older Persons' potentially inappropriate Prescriptions): application to acutely ill elderly patients and comparison with Beers' criteria. Age Ageing 2008; 37: 673–9.
16. Barry P, Gallagher P, Ryan C, O'Mahony D. START (Screening Tool to Alert doctors to Right Treatment). An evidence-based screening tool to detect prescribing omissions in elderly patients. Age Ageing 2007; 36: 628–31.
17. Cockcroft DW, Gault MH. Prediction of creatinine clearance from serum creatinine. Nephron 1976; 16: 31–41.
18. Levey AS, Bosch JP, Lewis JB *et al.* Modification of Diet in Renal Disease Study Group. A more accurate method to assess glomerular filtration rate from serum creatinine: a new prediction equation. Ann Intern Med 1999; 130: 461–70.
19. Levey AS, Green T, Kusek JW *et al.* A simplified equation to predict glomerular filtration rate from serum creatinine. J Am Soc Nephrol 2000; 11(Suppl): 155A.
20. Levey AS, Coresh J, Greene T *et al.* Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. Ann Intern Med 2006; 145: 247–54.
21. Spruill WJ, Wade WE, Cobb HH. Comparison of estimated glomerular filtration rate with estimated creatinine clearance in the dosing of drugs requiring adjustments in elderly patients with declining renal function. Am J Geriatr Pharmacother 2008; 6: 153–60.
22. Lamb EJ, Webb MC, O'Riordan SE. Using the modification of diet in renal disease (MDRD) and Cockcroft and Gault equations to estimate glomerular filtration rate (GFR) in older people. Age Ageing 2007; 36: 689–92.
23. Landis JR, Kock GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–74.
24. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990: 43: 551–8.

doi: 10.1093/ageing/afp058
Published electronically 12 May 2009

Vitamin D supplementation and type 2 diabetes: a substudy of a randomised placebo-controlled trial in older people (RECORD trial, ISRCTN 51647438)

SIR—Studies in animals show that vitamin D deficiency is associated with impaired insulin sensitivity, and that insulin